**INFM 600 – Info Seeking**

**Data Set 1:**

Contains 1,000,209 anonymous ratings of approximately 3,900 movies

made by 6,040 MovieLens users who joined MovieLens in 2000.

**Data Citation:**

================================================================================

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History

and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4,

Article 19 (December 2015), 19 pages. DOI=http://dx.doi.org/10.1145/2827872

URL for data: [http://grouplens.org/datasets/movielens/1m/](http://grouplens.org/datasets/movielens/1m/%20)

**Usage License:**

================================================================================

Neither the University of Minnesota nor any of the researchers

involved can guarantee the correctness of the data, its suitability

for any particular purpose, or the validity of results based on the

use of the data set. The data set may be used for any research

purposes under the following conditions:

\* The user may not state or imply any endorsement from the

University of Minnesota or the GroupLens Research Group.

\* The user must acknowledge the use of the data set in

publications resulting from the use of the data set

(see below for citation information).

\* The user may not redistribute the data without separate

permission.

\* The user may not use this information for any commercial or

revenue-bearing purposes without first obtaining permission

from a faculty member of the GroupLens Research Project at the

University of Minnesota.

**Description:**

This data set seems interesting not just because it contains ratings of Movies belonging to different genres but more importantly because these ratings have been given by a wide a range of users (6,040 users belonging to a wide range of age groups and occupations). What makes it interesting is that this data can be analyzed to dig into what kind of people are inclined to watching what kinds of movies. People tend to stereotype a person’s interests based on his/her age or occupation. But since these are anonymous ratings collected from a wide user base, there may be a few surprises about people’s movie preferences.

**Potential data users and decision-makers:**

A Movie-recommending website or application developer could use this data as a source to identify what movies to recommend based on people’s profiles. The data could also be of use to companies like Netflix to see if they can make some additions to their featured content based on these observations.

**Three questions this data might help to answer:**

1. How much does a person’s occupation influence his choice of movies?
2. Is there a pattern in the kind of movies a person tends to start liking as he/she grows up?
3. What is common in the movies that have been universally disliked by the audience?

**Data Set 2:**

Contains anonymous data of 618 subjects who were part of a Myopia Study to determine the factors that may potentially be responsible for the occurrence of Myopia among children.

**Data Citation:**

================================================================================

SOURCE: Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic

Regression: Third Edition. DOI= http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470582472.html

URL for data: <https://www.umass.edu/statdata/statdata/stat-logistic.html>

**Usage License:**

================================================================================

These data are copyrighted by John Wiley & Sons Inc. and must be acknowledged and

used accordingly.

**Description:**

This data set is interesting because it may help medical experts study the factors causing the occurrence of Myopia in children. The data has a variety of attributes ranging from the subject’s physical attributes (such as Spherical Equivalent Refraction, Axial Length, Lens Thickness etc.) to values such as the number of hours spent by the subject engaging in activities like sports, reading, watching TV etc.

The data therefore could be of great help in trying to understand what factors are most influential in the occurrence of Myopia and if there are ways to prevent it. The data set also contains data about whether the subject’s mother or father was myopic which in turn can be used to research the percentage of subjects that inherit the condition from their parents.

**Potential data users and decision-makers:**

The data could prove extremely beneficial to medical experts researching conditions like Myopia because it may provide interesting insights into the potential causes and prevention.

**Three questions this data might help to answer:**

1. Is myopia a hereditary condition?
2. What are the most common causes of myopia?
3. Is the condition influenced by the subject’s engagement in certain activities and can it be controlled?

**Data Set 3:**

This data set contains GPS trajectories collected from an android app called “*Go!Track*”.

**Data Citation:**

================================================================================

There are some papers that used this dataset. The first was presented in BRACIS (4th Brazilian Conference on Intelligent System).   
The paper was accepted but not published yet.   
The other paper was submitted to AT&T (IJCAI-16).   
The last paper is about the dataset.   
  
Publications:   
1 - CRUZ, M. O.; MACEDO, H.; GUIMARÃ£ES, A. P. Grouping similar trajectories for   
carpooling purposes. In: Brazilian Conference on Intelligent Systems. [S.l.: s.n.], 2015. p.   
234â€“239. ISBN 9781509000166.

URL for data: <https://archive.ics.uci.edu/ml/datasets/GPS+Trajectories>

**Terms of Use:** (No formal License details found for this data set)

================================================================================

If you publish material based on databases obtained from this repository, then, in your acknowledgements, please note the assistance you received by using this repository. This will help others to obtain the same data sets and replicate your experiments.

**Description:**

This data set has been collected from an android app called “*Go!Track*” and provides information about the GPS trajectories that users traversed in cars or buses. The data set contains detailed information such as the speed at which the user traveled, distance covered and time taken to cover a given track.

The interesting part about this data set is that each track covered by the user is uniquely identified by a ‘track id’ and is rated by the user based on traffic and weather conditions on that route. The data set also contains information about buses available on the traversed routes along with bus ratings as well. This data can therefore be used to identify the best routes possible between any given points and to check whether there is any good bus service available on it.

**Potential data users and decision-makers:**

The potential users for this data set could be android application developers who could use this data to recommend routes and bus services to daily commuters. Since the data set provides ratings for routes, buses and weather, this could be a good indicator for authorities to identify which areas are rated the worst and what can be done to improve the situation.

**Three questions this data might help to answer:**

1. What are the factors that influence a good or bad user rating for any route?
2. Which is the fastest route from point A to point B and is there a bus service available on the route?
3. Does the use of car or bus influence the rating for a route?